ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

Dynamic Virtual Machine Scheduling for Resource Sharing In the Cloud Environment

Karthika.M[1]

P.G Student, Department of Computer Science and Engineering, V.S.B Engineering College, Tamil Nadu[1]

ABSTRACT: Resource allocation and job scheduling are the core functions of cloud computing. These functions are based on adequate information of available resources. Timely acquiring dynamic resource status information is of great importance in ensuring overall performance of cloud computing. A cloud system for analyzing performance, removing bottleneck, detecting fault, and maintaining dynamic load balancing, thus, to help cloud users obtain desired computing results by efficiently utilizing system resources in terms of minimized cost, maximized performance between cost and performance. To reduce overhead, the aim of designing a dynamic resource allocation and prediction system is to achieve seamless fusion between cloud technologies and efficient resource scheduling and prediction strategies. This work aims at building a distributed system for cloud resource scheduling and prediction. In this project, we present the design and evaluation of system architecture for cloud resource scheduling and prediction. We discuss the key effects for system implementation, including virtual machine learning-based methodologies for modeling and optimization of resource prediction models. Evaluations are executed on a prototype system. Our experimental results suggest that the efficiency and accuracy of our system meet the demand of online system for dynamic resource utilization and prediction.

KEYWORDS- cloud computing, dynamic load balancing, prediction, resource scheduling.

I. INTRODUCTION

A Cloud is a type of parallel and distributed system which consists of a collection of interconnected and virtualized computers .The computing resources can be allocated dynamically upon the requirements and preferences of user. The consumers may access applications and data of the Cloud from anywhere at any time, it is hard for the cloud service providers to allocate the cloud resources dynamically and efficiently [1]. Physical resource are Computer, Processor, disk, database, network, Bandwidth, scientific instruments and the logical resources are Execution, monitoring, communicate application and etc.

In cloud computing, Resource Allocation (RA) is the process of assigning available resources to the needed cloud applications over the internet. Resource allocation starves services if the allocation is not managed precisely. Resource provisioning solves that problem by allowing the service providers to manage the resources for each individual module. Resource Allocation Strategy (RAS) is all about integrating cloud provider activities for utilizing and allocating scarce resources within the limit of cloud environment so as to meet the needs of the cloud application. It requires the type and amount of resources needed by each application in order to complete a user job. The order and time of allocation of resources are also an input for an optimal RAS. An optimal RAS should avoid the following criteria as follows:

a) **Resource contention** situation arises when two applications try to access the same resource at the same time.
b) S**carcity of resources** arises when there are limited resources.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

c) **Resource fragmentation** situation arises when the resources are isolated. [There will be enough resources but not able to allocate to the needed application.]

d) **Over-provisioning** of resources arises when the application gets

[

Today a growing number of companies have to process huge amounts of data in a cost-efficient manner. In order to simplify the development of distributed applications on top of such architectures, many of real time companies have also built customized data processing frameworks. Today's processing frameworks typically assume the resources they manage consist of a static set of homogeneous compute nodes. However designed to deal with independent nodes failures, they examine the number of available machines to be constant, mostly when scheduling the processing job's execution. Hence, the compute resources available in a cloud are highly dynamic and possibly heterogeneous.Regarding to parallel data processing, this flexibility guides to a variety of new possibilities, especially for scheduling data processing jobs. This new model allows allocating compute resources dynamically and just for the time they are required in the processing workflow.

Even if it was possible to determine the underlying network hierarchy in a cloud and use it for topology-aware scheduling, the obtained information would not perfectly remain valid for the entire processing time. VMs may be moved for administrative purposes between different locations inside the data center without any notification, providing any previous knowledge of the relevant network infrastructure obsolete. As a result, the only way to assure locality between tasks of a processing job is currently to execute these tasks on the same VM in the cloud. This may involve assigning fewer, but more powerful VMs with multiple CPU cores.

For, e.g., a framework exploiting the possibilities of a cloud could start with a single VM which analyzes an incoming job and then advises the cloud to directly start the required VMs according to the job's processing phases. After every phase, the machines could be released and no longer contribute to the overall cost for the processing job. Facilitating such use cases exploits some requirements on the design of a processing framework and the way its jobs are described. First, the scheduler of that framework must become aware of the cloud environment a job should be executed in. It must know about the several types of available providers as well as their cost and be able to allocate or destroy them on behalf of the cloud customer. Second, the paradigm used to describe jobs must be powerful enough to express dependencies between the different tasks the job consists of.

II REVIEW OF RELATED WORKS

Cloud computing removes the limitation that exist in traditional shared computing environment, and becomes a leading trend in distributed computing system. It accumulates heterogeneous resources distributed across Internet, regardless of differences between resources such as platform, hardware, software, architecture, language, and geographical location. Such resources, which include computing, storage, data, communication bandwidth resources and other resources, are combined dynamically to form high performance computing capability of solving problems in large-scale applications. Dynamically sharing resources raising the resource contention. The challenging problems are deciding the destination nodes where the tasks of cloud application are to be executed. From the perspective of system architecture, dynamic resource allocation and scheduling are the most crucial functions of cloud computing.

However, the processing frameworks which are currently used have been designed for static, homogeneous cluster setups and disregard the cloud nature. Consequently, the allocated compute resources may be inadequate for big parts of the submitted job and unnecessarily increase processing time and cost. Resource allocation alone, however, can only support instantaneous resource information acquisition. It cannot generalize the dynamic variation of resources. Resource state prediction is inevitable to fill this gap. Typical previous prediction systems, can provide both allocation and prediction functions. However, dynamic features of cloud resources were not taken into consideration in these design frameworks.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

Drawbacks

- This existing works focuses on the scheduling problem in dynamic cloud resource allocation.
- Many previous research projects focused on optimizing traditional performance metrics, like system utilization, client incentive and application response time in controlled cloud. They did not consider dynamic resource allocation based on provider load.
- Particular tasks of a processing job can be assigned to different types of virtual machines which are automatically instantiated and terminated during the job execution.
- Dynamically sharing resources increases resource contention.

In a cloud this topology information i.e. provider information is typically exposed to the customer. Since the nodes involved in processing a data intensive job often have to transfer tremendous amounts of data through the network, this drawback is particularly severe; parts of the network may become congested while others are essentially unutilized. Even if it was possible to determine the underlying network hierarchy in a cloud and use it for topology- aware scheduling, the obtained information would not necessarily remain valid for the entire processing time. Based on the challenges and opportunities outlined in the previous section we have designed our proposed work, a new data processing framework for cloud environments. It takes up many ideas of previous processing frameworks but refines them to better match the dynamic and opaque nature of a cloud. Resource monitoring alone, however, can only support instantaneous resource information acquisition. It cannot generalize the dynamic variation of resources. Resource state prediction is inevitable to fill this gap.

For example, it considers aspects like the input/output cardinalities of the first-order functions which are helpful to deduce reasonable degrees of parallelization. The main challenge, phrased as a scheduling problem, is to schedule jobs of consumers to resources of providers to optimize incentives for both parties with respect to parallel data processing. Considering the heterogenous and dynamic characteristics of computing cloud, resource monitoring and prediction.

III SYSTEM OVERVIEW

In this paper, we present the design and evaluation of system architecture for cloud resource scheduling and prediction. The challenge is to develop a cloud scheduling scheme that enables prediction based resource allocation by middleware to make autonomous decisions while producing a desirable emergent property in the cloud system; that is, the two system wide objectives are achieved simultaneously. We discuss the key issues for system implementation, including machine learning based methodologies for modeling and optimization of resource prediction models.

There are mainly two mechanisms for acquiring information of cloud resources: cloud resource monitoring and cloud resource prediction. Cloud resource state monitoring cares about the running state, distribution, and system load in cloud system by means of monitoring strategies. Cloud resource state prediction focuses on the variation trend and running track of resources in cloud system by means of modeling and analyzing each provider's load i.e CPU usage. Periodic updation by monitoring and future variation generated by prediction are combined together to feed cloud system for analyzing performance, eliminating diagnosing fault, and maintaining dynamic load balancing, thus to help cloud users obtain desired computing results by efficiently utilizing system resources in terms of minimized cost, maximized performance or tradeoffs between cost and performance. To reduce overhead, the goal of

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

designing a cloud resource monitoring and prediction system is to achieve seamless fusion between cloud technologies and efficient resource monitoring and prediction strategies.

The following aspects play an important role in designing cloud infrastructure:

Client-server cloud computing or networking is a distributed application architecture that partition tasks or workloads between resource providers and service requesters, called clients. Virtual machine or middleware is the system maintains the details of the client and the providers. Virtual machine maintains the details of each connected resource providers which is used for job scheduling, resource provider selection and for task splitting.

Providers are the process to satisfying the client request. Providers are varied under system configuration, workload and performance. Providers needs to register their profile with the virtual machine, thus virtual machine can able to allocate the job under client request. Virtual machine gets the short listed providers system current CPU usage which is used for prediction to allocate the task.

Virtual machine collects the job request from client and assigns the job to the provider which providing the respected service to the client. That provider has been selected based on the CPU usage who having the less CPU utilization. Long job request can be manipulated in an efficient manner by using parallel processing technique.

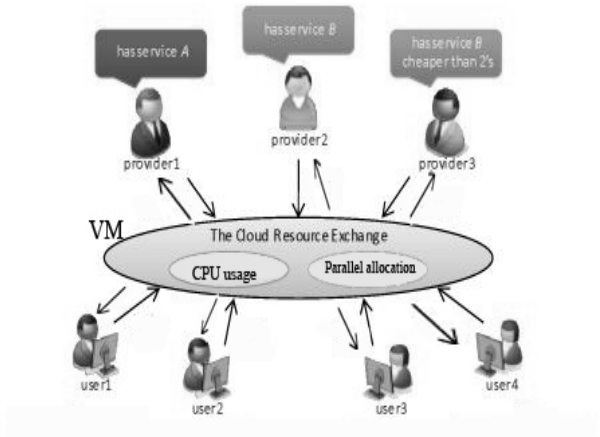

Fig 1.Architecture Diagram

A Parallel processing technique

By using this technique long job can be splitted into sub tasks and select the respective providers through their prediction method based on requested resource provider's availability and virtual machine allocate the each subtask to the selected providers in parallel. After task allocation virtual machine monitor the process execution in each provider. Then through mapping send the response to client.

Eg task client request: Prime number calculation from 1 to 100000

Virtual machine predict the providers (eg. 4 provides) based on system load and split the task as 1 to 25000, 25000 to 50000, 50000 to 75000 and 75000 to 100000 and allocate to each providers in parallel instead of assign to one resource provider. So through our approach resource provider's utilization and the performance must be improved as well as client's

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

response time must be reduced.After completion of sub task, server maps into task and deliver the task output to the respective client. Thus through our approach overall cloud utilization and performance is measured perfectly and profitably.

B.Implementation

The implementation procedures are as follows:

1. Server or virtual machine plays a role between client and the resource providers. Before client job request, server gets the resource providers details like resource name and related details with its current system connectivity which is used for prediction based dynamic resource allocation.

2. Client submits job to the middleware instead of provider selection done by client, then middleware or server or virtual machine schedule the job through predicting system load (which as system CPU usage) of the short listed resource providers. Based on system low CPU usage of multiple providers, server allocates the job to the provider whose load i.e. current CPU usage is lower.

3. Then server collects the result from the providers and sends to the respective client. Thus client waiting time or response time must be reduce and also each provider utilization time and energy must be less so overall performance must be better than the existing work.

IV CONCLUSION

Design and evaluation of system architecture for cloud resource scheduling and prediction achieve the following aspescts:

- Dynamic resource allocation executed in cloud with good performance metric values.
- Improve the overall utilization of server resources.
- The performance evaluation gives a first impression on how the ability to assign specific virtual machine types to specific tasks of a processing job, as well as the possibility to automatically allocate/deallocate virtual machines in the course of a job execution, can help to improve the overall resource utilization and, consequently, reduce the processing cost.

REFERENCES

[1] `L. Siegele, "Let It Rise: A Special Report on Corporate IT," The Economist, vol. 389, pp. 3-16, Oct. 2008.

[2] M. Nelson, B.-H. Lim, and G. Hutchins, "Fast Transparent Migration for Virtual Machines," Proc. USENIX Ann. Technical Conf., 2005.

[3] T. Wood, P. Shenoy, A. Venkataramani, and M. Yousif, "Black-Box and Gray-Box Strategies for Virtual Machine Migration," Proc. Symp.Networked Systems Design and Implementation (NSDI '07), Apr. 2007.

[4] C.A. Waldspurger, "Memory Resource Management in VMware ESX Server," Proc. Symp. Operating Systems Design and Implementa- tion (OSDI'02), Aug. 2002.

[5] '09 G. Chen, H. Wenbo, J. Liu, S. Nath, L. Rigas, L. Xiao, and F. Zhao, "Energy-Aware Server Provisioning and Load Dispatching for Connection-Intensive Internet Services," Proc. USENIX Symp. Networked Systems Design and Implementation (NSDI '08), Apr. 2008.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

[6] Zhen Xiao, Senior Member, IEEE, Weijia Song, and Qi Chen," Dynamic Resource Allocation Using Virtua Machines for Cloud Computing Environment" IEEE transactions on parallel and distributed systems, vol. 24, no. 6, june 2013

[7] P. Padala, K.-Y. Hou, K.G. Shin, X. Zhu, M. Uysal,Z. Wang, S.Singhal, and A. Merchant, "Automated Control of Multiple Virtualized Resources," Proc.ACM European conf. Comp System (EuroSys),

[8] T. Das, P. Padala, V.N. Padmanabhan, R. Ramjee, and K.G. Shin, "Litegreen: Saving Energy in Networked Desktops Using Virtua-lization," Proc.USENIX Ann.Technical Conf., 2010

[9] Jiyani et al.: Adaptive resource allocation for preemptable jobs in cloud systems (IEEE, 2010), pp.31-36.

[10] K.H Kim et al. Power-aware provisioning of cloud resources for real time services. In international workshop on Middlleware for grids and clouds and escience, pages 1-6, 2009.

[11] Karthik Kumar et al.: Resource Allocation for real time tasks using cloud computing (IEEE, 2011), pp.

[12] Keahey et al., "sky Computing",Intenet computing, IEEE,vol 13,no.5,pp43-51,sept-Oct2009.